A JOURNEY THROUGH FLOW CYTOMETRIC IMMUNOFLUORESCENCE ANALYSES

- Finding accurate and robust algorithms
that estimate positive fraction distributions -

by

C. Bruce Bagwell, MD, Ph.D.
Verity Software House, Inc.

Published: Clinical Immunology Newsletter, Vol. 16, No. 3, 1996.

Over the last decade there have been a number of proposed computer algorithms to estimate positive cells in immunofluorescence histograms (1,2,3). A simple description of the problem being solved is

> ***"Find the most probable proportions of negative and positive staining distributions in a test histogram given only the shape and position of the negative distribution."***

The appropriate decomposition of a test histogram into negative and positive distributions is show in Figure 1.

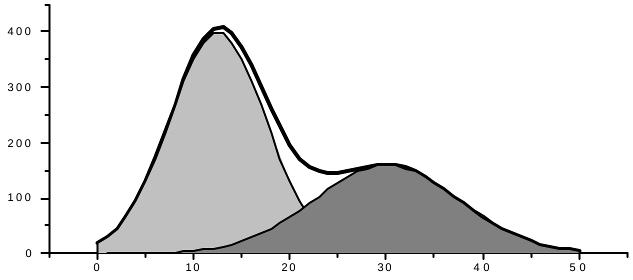

Figure 1: Simple decomposition of a test distribution into negative and positive distributions.

Although at first glance this problem seems trivial, an exact solution may not ever be possible given the limited information available to the analysis routines. Consider the following equivalent solution to the problem shown in Figure 2:

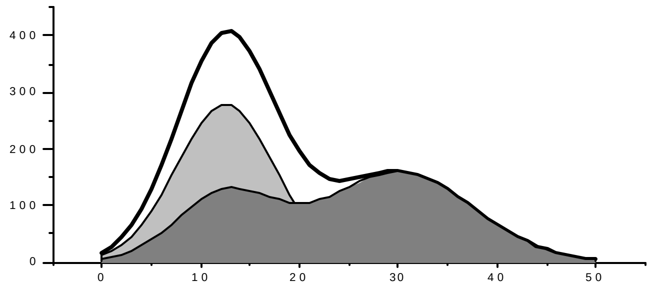

Figure 2: Equivalent decomposition of a test distribution into negative and positive distributions.

The critical missing piece is the specification for the shape and position of the positive distribution. Because of this limitation, all proposed analysis methods that make no assumptions about the positive distribution are approximations, not exact solutions.

In our journey through some of these approximations, we will encounter a simple approach with egregious sources of errors yielding surprisingly accurate estimates and a more refined approach yielding less accurate results. We will see a routine that is really a 60 year old statistical test in disguise which, unbeknownst to its Russian inventor, works because it estimates an additional unknown distribution such as the positive distribution. We will discover that many of the seemingly diverse proposed methods can be described by a more general theory which directly leads to more accurate and robust analytical methods.

Before we begin, we need to discuss how this journey will be organized. Each method will be described by mathematical formulae and evaluated and compared by means of an arbitrary example (see Fig. 3). The estimated positive fraction will be shown with its corresponding error. At the end, the methods will be compared by a more thorough simulator that produces histograms of various proportions and shape (8).

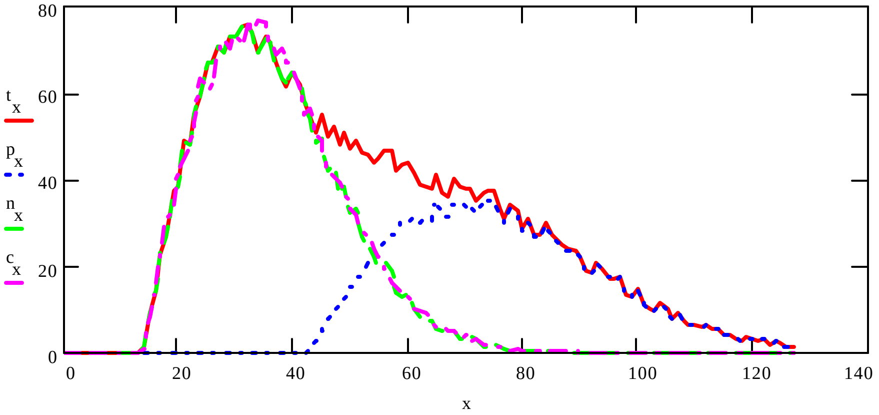

Figure 3: Example data set (pos=0.4).

The formulae will be composed of the symbols summarized in Table 1.

Symbol	Definition
U	upper channel in all histograms (e.g. 127)
L	lower boundary, depends on method
x	channel value, ranges between 0 and U
c_x	control histogram evaluated at x
t_x	test histogram evaluated at x
p_x	positive histogram evaluated at x
n_x	negatives in test histogram evaluated at x
C_x	cumulative control histogram, normalized to range from 0 to 1 at x
T_x	cumulative test histogram, normalized to range from 0 to 1 at x
D_x	difference between C_x and T_x
P_x	cumulative positive histogram at x
x_d	x channel with maximum absolute difference (i.e. max(D))
x_{d2}	Dmax for normalized cumulative control and test in interval 0..x_d
c_t	total number of events in the control histogram
t_t	total number of events in the test histogram
pos	actual positive fraction
pos_m	estimated positive fraction for method “m”
ε	error, defined as (calculated-actual)*100/actual.

Table 1: Summary of all mathematical symbol definitions.

Scenic View 1: Integration Method (pos_i)

The lower boundary of integration, L, is normally found by finding the element in **C**$_x$ that is equal to or just less than some defined fraction, e.g. 0.95.

$$pos_i := \frac{\sum_{x\,L}^{U} t_x}{t_t} \qquad pos_i = 0.376 \qquad \varepsilon_i = -5.967 \qquad \text{[Eq. I-1]}$$

Integration has two main sources of error; false positives and negatives. For this example, its error is quite reasonable given the size of its individual sources of error.

Scenic View 2: Enhanced Integration Method (pos_{ei}):

The integration routine can theoretically be improved by calculating the fraction of negatives that are in the positive integration region and subtracting that fraction from the pos_i estimate. The formula is easily derived as

$$pos_{ei} := pos_i - \left(1 - C_L\right) \cdot \frac{1 - pos_i}{C_L} \qquad pos_{ei} = 0.346 \qquad \varepsilon_{ei} = -13.338 \qquad \text{[Eq. I-2]}$$

Surprisingly, the error associated with this improvement is higher than the simpler integration method.

The error increases because the two sources of error for the Integration method partially cancel each other out. Thus, by removing only the false positive source of error, the Enhanced Integration Method error is increased compared to the Integration Method. The integration method works best with symmetric distributions and deteriorates rapidly as the distributions become skewed. The sensitivity to distribution shape precludes this method as a robust positive fraction estimator.

Scenic View 3: Dmax Method (pos_d):

In the 1930's Andrei Nikolaevich Kolmogorov and later N Smirnov developed a statistic that quantified the difference between two frequency histograms with the following design characteristics: 1) no required assumptions about error distributions, 2) easy manual calculation, and 3) robust with noisy data (4,5).

The statistic rapidly became one of the most widely used nonparametric tests for histogram comparison and eventually was labeled with its author's names, Kolmogorov and Smirnov. The Kolmogorov-Smirnov or KS statistic is the maximum absolute difference, or Dmax, between two cumulative probability distributions. The test histogram cumulative distribution is computed as shown below:

$\mathbf{T}_0 = t_0$
$\mathbf{T}_x = \mathbf{T}_{x-1} + t_x$ for x>0
$\mathbf{T}_x = \mathbf{T}_x/t_t$ for all x.

A probability distribution is a histogram normalized with an area of one. Presumably, Kolmogorov and Smirnov selected cumulative distributions because their successive summations have a powerful smoothing effect on noisy data. Figure 4 shows the example test and control cumulative probability distributions and the maximum absolute difference, Dmax.

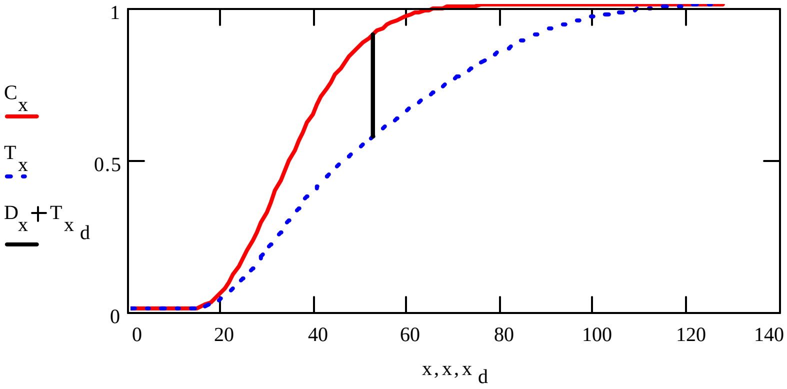

Figure 4: Control and Test cumulative distributions and their maximum difference, Dmax.

The KS test was first suggested as a histogram comparison test for flow cytometry in 1977 (6) and two years later it was used to detect B-cell clonality in bone marrow derived Kappa and Lambda stained histograms (7). The first hint that Dmax was more than just a nonparametric statistic came from the clonality data. The Kappa-Lambda sensitivity was estimated to be 10% blasts and its corresponding Dmax was 0.10. As will be seen, this identity was no coincidence.

In 1988 a method called cumulative subtraction (CS,3) was proposed to estimate percent positives. On close inspection, the CS algorithm is equivalent to the venerable KS Dmax statistic.

$D_x := C_x - T_x$
$pos_d := \max(D) \qquad pos_d = 0.336 \qquad \varepsilon_d = -15.971$ [Eq. D-1]

The error associated with this estimate is relatively high and demonstrates a strong propensity to underestimate the true positive fraction.

Scenic Overlook

In order to understand how Dmax approximates the positive fraction and to improve upon its performance, we developed a more general positive fraction-cumulative distribution equation.

$$\mathrm{POS}(x) := \frac{D_x + P_x}{C_x} \qquad \text{[Eq. D-2]}$$

A plot of POS(x) versus x (see Fig. 5) demonstrates that for all x>>0, Equation D-2 estimates the positive fraction. The plot also shows the smoothing effect of cumulative histograms. By channel 40 the random effects of noisy data have almost been eliminated. With noiseless data, the actual positive fraction and POS(x) are identical for all x>0 with P_x values>0.

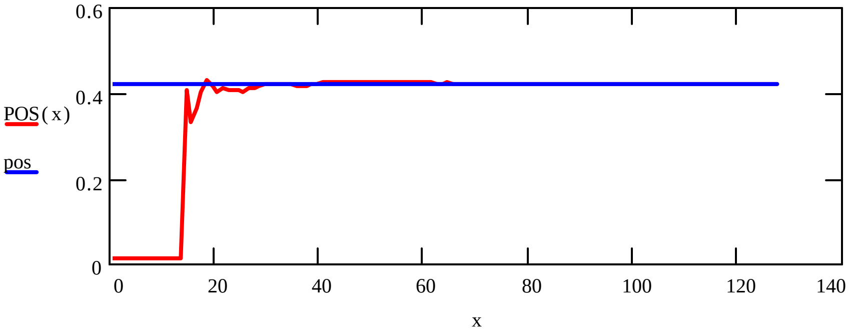

Figure 5: Demonstration of the general positive fraction-cumulative histogram theory for predicting positive fraction.

Inspection of Equation D-2 reveals that as $\mathbf{D}_x$ approaches its maximum value, $\mathbf{P}_x$ approaches 0 and $\mathbf{C}_x$ approaches 1; thus, max($\mathbf{D}_x$) or Dmax approximates the positive fraction. It is also clear from the formula that Dmax will always underestimate the true positive fraction.

The KS test Dmax is more general than just a positive fraction estimator. The absolute value operator makes it an estimator of an additional distribution in either of the histograms, thus making it a perfect test for B-cell clonality.

Scenic View 4: Enhanced Dmax Method ($\mathrm{pos}_{\mathrm{ed}}$):

Equation D-2 suggests an easy way to improve the Dmax as a predictor of positive fraction. Since $\mathbf{C}_x$ is known, we can divide the Dmax by $\mathbf{C}_{xd}$ to find an improved estimate.

$$\mathrm{pos}_{ed} := \frac{D_{x_d}}{C_{x_d}} \qquad \mathrm{pos}_{ed} = 0.383 \qquad \varepsilon_{ed} = -6.557 \qquad \text{[Eq. D-3]}$$

Scenic View 5: Normalized Subtraction Method (pos_{ns}):

In 1981 a seemingly unrelated method for estimating positive fraction was proposed (1,2). In normalized subtraction an amplification factor, k, multiplies the control histogram such that within a certain match interval the two histograms, control and test, have equivalent areas. The normalized control histogram is then subtracted from the test histogram and the positive difference is the estimated positive distribution.

If the amplification factor, k, is evaluated over the interval $0..x_d$, and the difference residue is summed regardless of sign, it can be shown that the estimated positive fraction is identical to the Enhanced Dmax technique.

$$k_{ns} := \frac{\sum_{i=0}^{x_d} t_i}{\sum_{i=0}^{x_d} c_i} \qquad \mathrm{pos}_{ns} := \frac{\sum_x \left(t_x - k_{ns} \cdot c_x\right)}{t_t} \qquad \begin{matrix} \mathrm{pos}_{ns} = 0.383 \\ \mathrm{pos}_{ed} = 0.383 \end{matrix}$$

Thus, Equation D-2 embraces the seemingly disparate methods of Cumulative Subtraction or Dmax, Enhanced Dmax, and Normalized Subtraction.

Scenic View 6: Enhanced Normalized Subtraction Method (pos_{ens}):

Equation D-2 also suggests that if we could somehow estimate $\mathbf{P}_x$, the estimation of positive fraction would further improve. Evaluating equation D-2 at the location of maximum difference, x_d, yields:

$$\mathrm{POS}\left(x_d\right) = \frac{D_{x_d} + P_{x_d}}{C_{x_d}}$$

The term P_{xd} defined as is the fraction of positives in the interval $0..x_d$. If we estimate this fraction by means of a second Dmax ($x=x_{d2}$) over the interval $0..x_d$ between normalized cumulative control and test histograms, we arrive at the formula

$$\mathrm{pos}_{\mathrm{ens}} := \frac{C_{x_d} - T_{x_d}}{C_{x_d}} + \frac{C_{x_{d2}} \cdot T_{x_d} - C_{x_d} \cdot T_{x_{d2}}}{\left(C_{x_d}\right)^2} \qquad \mathrm{pos}_{\mathrm{ens}} = 0.397 \quad \varepsilon_{\mathrm{ens}} = -0.545 \qquad \text{[ENS-1]}$$

Scenic Overview:

If we synthesize thousands of diverse histograms representing a wide variety of immunofluorescence histograms (8), we can better appreciate the performance of the Dmax, Enhanced Dmax, and Enhanced Normalized Subtraction methods.

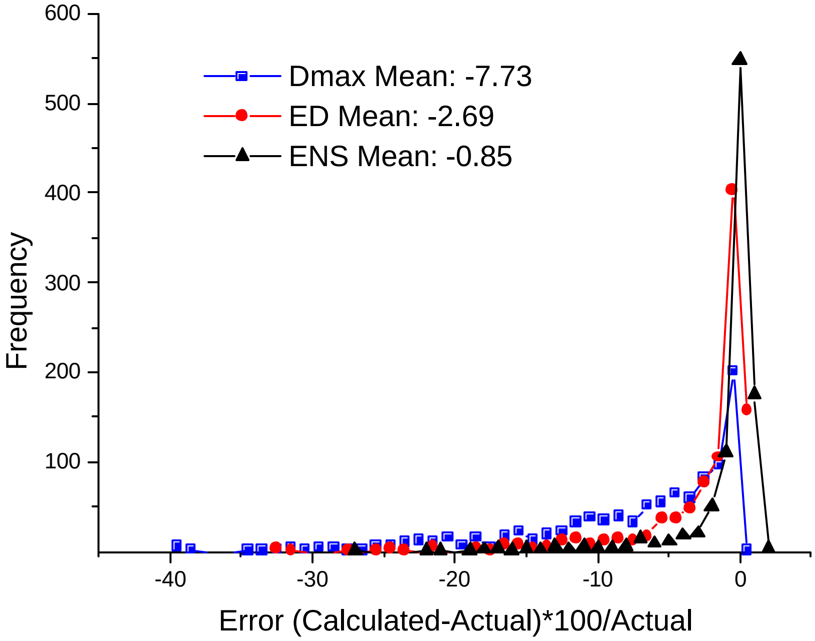

Figure 6: Error distributions for D, ED, and ENS methods.

The ENS method has a mean error of -0.85, ED was -2.69, and D was -7.73. Head-to-head comparisons show that the ENS method will outperform the D method at a rate of at least 100 to 1 and the ED method at 3 to 1.

Summary of Route:

In our journey, we found the simple approach of integration is surprisingly accurate given its large false positive and negative errors. With symmetric negative and positive distributions, these errors largely cancel each other. Attempts to improve the estimate by accounting for the false positive error, inevitably leads to less accurate estimates.

The early KS Dmax statistic is really an estimator for the positive fraction. A general formula that embodies this approximation also predicts better ways of estimating positive fractions. The Enhanced Normalized Subtraction method was found to have the best analysis characteristics of all the methods described.

Reference:

1. Hoffman, R. Simple analysis of immunofluorescence histograms. Abstract 61. VIII Converence on Analytical Cytology and Cytometry, Wentworth-by-the-Sea, New Hampshire: May 19-25. Cytometry 1981.
2. Bagwell CB: IMMUNO program, EASY2 Software, Coulter Electronics, Inc., 1981.
3. Overton RW. Modified histogram subtraction technique for analysis of flow cytometry data. Cytometry 1988;9:619-626.
4. Kolmogorov A. Sulla determinazione empirica di una legge di distribuzione. Giornalle dell Instituto Italiano degli Attuari 1933; 4: 1-11.
5. Smirnov N. On the estimation of the discrepancy between empirical curves of distribution for two independent samples (In Russia). Bull Moscow Univ. Intern Ser Math, 1939; 2:3-16.
6. Young IT. Proof without prejudice: Use of Komogorov-Smirnov test for the analysis of histograms from flow systems and other sources. J Histochem Cytochem 1977;25:935-941.
7. Ault KA. Detection of small numbers of monoclonal B lymphocytes in the blood of patients with lymphoma. N Engl J Med 1979;300:1401-1405.
8. Simulation specifics: Microsoft Access was used to generate and database the analysis results. Negatives ranged from 200 to 1000, positives from 200 to 1000. A Weibull was used to model both distributions.

$$\mathrm{weibull}(x, c, w, s, ph) := \left|\begin{array}{l} f2 \leftarrow \dfrac{s-1}{s} \\[2ex] f1 \leftarrow \dfrac{(x-c)}{w} + (f2)^{\frac{1}{s}} \\[2ex] f1 \leftarrow 0 \quad \text{if} \quad f1 < 0 \\[2ex] y \leftarrow ph \cdot (f2)^{-f2} \cdot (f1)^{s-1} \cdot e^{-(f1)^{s} + f2} \\ 0 \quad \text{if} \quad y < 0 \end{array}\right.$$

For the control population c, w, and s ranged from 20-40, 10-40, and 2-2.5 and the positive population ranged from 50-80, 10-40, and 2-2.5 respectively. Statistically noise was added appropriately to all histograms.